# AUTOMATIC EXTRACTION OF GEOMETRIC LIP FEATURES WITH APPLICATION TO MULTI-MODAL SPEAKER IDENTIFICATION

*Ivana Arsic, Roger Vilagut and Jean-Philippe Thiran*

Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email:{Ivana.Arsic, Roger.Vilagut, JP.Thiran}@epfl.ch
http://itswww.epfl.ch

## ABSTRACT

In this paper we consider the problem of automatic extraction of the geometric lip features for the purposes of multi-modal speaker identification. The use of visual information from the mouth region can be of great importance for improving the speaker identification system performance in noisy conditions. We propose a novel method for automated lip features extraction that utilizes color space transformation and a fuzzy-based c-means clustering technique. Using the obtained visual cues closed-set audio-visual speaker identification experiments are performed on the CUAVE database, [1] showing promising results.

## 1. INTRODUCTION

An exhaustive research done in the area of joint audio-visual speech processing over the last decade shows that the usage of visual modality can be beneficial for improving the overall system performance, especially in the cases when the environmental conditions are changed. The inclusion of visual modality together with audio have been already used in various applications such as: audio-visual speech recognition [2], bimodal speaker recognition [3], speaker localization, etc.

The main problem when trying to benefit from speech and speaker dependent information conveyed in the lip movement, is the extraction of the suitable lip features. Moreover, the applied method should be robust and accurate, and if possible accomplished in an automatic manner. Various techniques were proposed by different researchers and an excellent overview of different methods can be found in [2].

In this paper we address the problem of automatic extraction of geometric lip features, namely four outer lip contour points from the observed lip region images. Further on, the obtained geometric lip descriptors are applied in the framework of audio-visual speaker identification on the CUAVE database [1], previously used mainly for audio-visual speech recognition. The recent work by Dean et al. [4], is the first showing that this audio-visual corpora can be equally used to test speaker identification performance using eigenlips as visual cues. This motivated us to test the possibility of using geometric based features for the same tasks.

Section 2 describes an algorithm for extraction of geometric lip features. In Section 3 we recall the mathematical background regarding the speaker modeling using Gaussian mixture models. The experimental framework and obtained results are presented in Section 4, followed by the conclusions in Section 5.

The work is supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

## 2. VISUAL FEATURE EXTRACTION

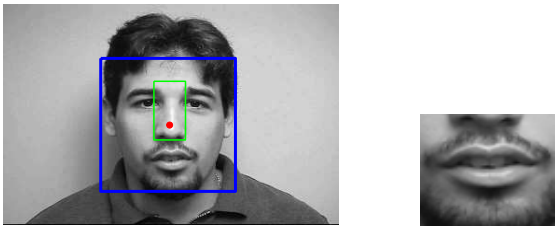
The extraction of visual features is performed in two stages. The first considers the lip Region-of-Interest (ROI) localization in each frame of the audio-visual sequence. In the second stage a lip segmentation technique based on fuzzy c-means clustering is applied on the lip ROI and features are extracted.

### 2.1. Lip ROI extraction

Prior to the lip Region-of-Interest (ROI) extraction, a face detection is performed using the algorithm described in [5]. As a result, the square bounding box and a center of the subject's face is found. Then we use the template matching of the nose region between consecutive frames in order to stabilize the detection process. Knowing the nose (face) center position, we use anthropometrical properties of the human face to define a region around the center of the mouth from the first frame. Subsequently, from each frame of size $360 \times 240$ of the observed video sequence the $60 \times 50$ pixels RGB color lip ROI is localized and extracted, as shown in Figure 1.

### 2.2. Color space transformation

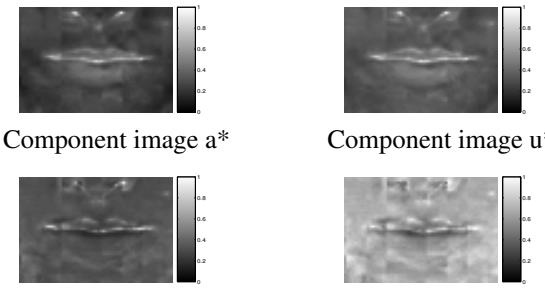
As a first step a color transformation from the RGB color space to CIELAB and CIELUV is performed on the extracted lip ROI. The main idea is to obtain perceptually uniform color spaces, as well as to increase robustness to the lip/skin color



(a) (b)

**Fig. 1**. Example of the face detection (a) and extracted lip ROI (b).

variations across speakers. After color transformation, each pixel in the image is represented by the five dimensional feature vector $(L^*, a^*, b^*, u^*, v^*)$ [6]. Since we want to have luminance-free conditions, the $L^*$ component is not taken into account. Thus, the size of the feature vector is reduced to four. The result of the color transformation of the mouth region is shown in Figure 2. Each separate color component is treated as an image and called a *component image*.



Component image a* Component image u*

Component image b* Component image v*

**Fig. 2**. The color component images after applying color transformation on the lip ROI.

### 2.3. Lip segmentation using fuzzy clustering

For segmenting the lip area from the extracted ROI we use the algorithm based on fuzzy c-means clustering, similar to the one presented in [7]. The lip features of interest are four outer lip contour points, namely lip corners and upper and lower lip middle points. The extraction procedure is performed through an iterative process in the following three steps, repeated until convergence: masks and centroids calculation, membership function calculation, and feature extraction and error correction.

### 2.4. Mask and centroids calculation

For each *component image* an initial set of masks $M_i$ is built using the color distribution information from the image histogram and morphological filtering. Color centroids are calculated by overlaying each component mask to the corresponding *component image*. Hence, we can calculate color centroids for lip region $v_{lip}$ and skin region, $v_{skin}$, as following:

$$\begin{aligned} v_{lip,i} &= \frac{1}{l_i} \sum x^*(r,c) \cdot M_i(r,c) \\ v_{skin,i} &= \frac{1}{m \cdot n - l_i} \sum x^*(r,c) \cdot (1 - M_i(r,c)) \end{aligned} \quad (1)$$

where $x^*(r,c) \in \{a^*, b^*, u^*, v^*\}$ represents a pixel in the row $r$ and column $c$ in the image of dimension $m \times n$. Parameter $l_i$ denotes the size of the mask, i.e. the number of pixels having the intensity equal to 1 for the lip region, and 0 for the skin region.

### 2.5. Membership function calculation

The next step is a membership function calculation and the objective is to maximize the dissimilarity between pixels from different clusters.

Let $d_i(r,c) \in \{d_{i,a}(r,c), d_{i,b}(r,c), d_{i,u}(r,c), d_{i,v}(r,c)\}$ stand for the Euclidean distance between the feature vector of each pixel $x(r,c)$ and the color centroids $v_{lip}$ and $v_{skin}$. We calculate the color dissimilarity $D_i(r,c)$ for lip and skin as:

$$\begin{aligned} D_{lip}(r,c) &= d^2_{lip,a}(r,c) + d^2_{lip,u}(r,c) \\ &+ \frac{1}{3}(d^2_{lip,b}(r,c) + d^2_{lip,v}(r,c)) \quad (2) \\ D_{skin}(r,c) &= d^2_{skin,a}(r,c) + d^2_{skin,u}(r,c) \\ &+ \frac{1}{3}(d^2_{skin,b}(r,c) + d^2_{skin,v}(r,c)) \quad (3) \end{aligned}$$

In equations 2 and 3 the experimentally obtained weighting parameter $\frac{1}{3}$ is introduced in order to decrease the influence of masks $M_b$ and $M_v$ since they can be of a poor quality. Finally, the membership map for the lip cluster is calculated as:

$$u_{lip}(r,c) = \frac{D_{skin}(r,c)}{D_{lip}(r,c) + D_{skin}(r,c)} \quad (4)$$

The resulting membership map of an example lip ROI at each iteration is shown in Figure 3 (a). In order to segment the lip area, the probability map is thresholded at 0.5 and the biggest connected area represents the lips. The undesired background effects can be eliminated by utilizing the lip symmetry. Segmented lip images after thresholding for different iteration steps are shown in Figure 3 (b).

### 2.6. Lip features extraction

After locating the lip area in the lip ROI image, we can estimate coordinates of the lip features of interest. The left and right lip corners are found from the $x$ and $y$ image intensity profiles. Then, the mouth center coordinates are calculated by finding the mean value of the distance between the lip corners. An orthogonal to the line connecting the lip corners is

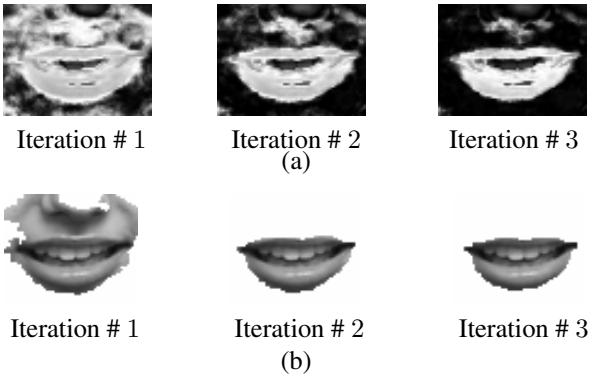



**Fig. 3**. Visualization of the probability map throughout the iterative process (a), and segmented lips (b).

found and used as a search line for the upper and lower outer contour central points.

Having the estimated key lip contour points' coordinates we can calculate the height and width of the mouth. Let $h_{k,j}$ and $w_{k,j}$ denote the height and width of the mouth, respectively for frame $k$ at iteration $j$ and $h_{k-1}$ and $w_{k-1}$ denote the same parameters for the last iteration of the previous frame. At the end of each iteration the estimated height and width are first compared with those of the previous frame, and after to those of the previous iteration. Convergence of the iterative process is defined by the following conditions:

$$\begin{aligned} |h_{k,j} - h_{k-1}| \leq \varepsilon_{h,k}, \quad |w_{k,j} - w_{k-1}| \leq \varepsilon_{w,k} \\ |h_{k,j} - h_{k,j-1}| \leq \varepsilon_{h,j}, \quad |w_{k,j} - w_{k,j-1}| \leq \varepsilon_{w,j}. \end{aligned} \quad (5)$$

The appropriate values for $\varepsilon_{h,k}$, $\varepsilon_{w,k}$, $\varepsilon_{h,j}$ and $\varepsilon_{w,j}$ are determined experimentally from an evaluation set of manually marked feature points for 100 frames per each speaker in a set of ten speakers. After calculating the maximum variations of height and width between consecutive frames, the values for $\varepsilon_{h,k}$ and $\varepsilon_{w,k}$ are set to be equal four pixels. When these conditions are not satisfied the iterative process starts again with the mask calculation.

In order to account for possible errors introduced when the lip corner coordinates are not well placed an additional error correction step is introduced at the end of each iteration. The maximum allowed pixel displacement between the consecutive frames is set to three pixels. If the decision threshold is exceeded, the pixel coordinates of the previous frame are kept. As a final step, temporal smoothing using a median filter is applied to the extracted set of lip coordinates. Some example image frames of different speakers and located lip points after running the algorithm are shown in Figure 4.

## 3. SPEAKER MODELING

The baseline of our speaker identification system is built using a state-of-the art Gaussian mixture model approach (GMM)

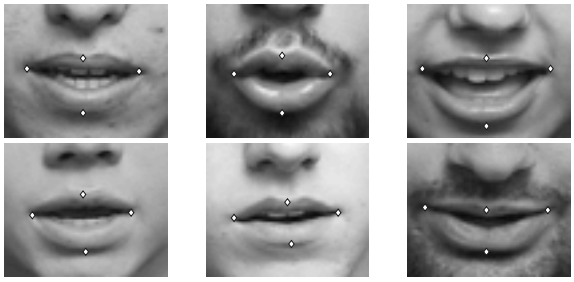


**Fig. 4**. Results of the lip feature extraction algorithm.

as in [8]. One GMM model is built for each speaker from the database, representing all the speech utterances of that speaker. Thus, a speaker $S$ is represented by a model $\lambda_S$:

$$\lambda_S = \{p_i^S, \mu_i^S, \Sigma_i^S\} \qquad with \quad i = 1, \ldots, M. \quad (6)$$

where $p_i^S$, $\mu_i^S$ and $\Sigma_i^S$ are the mixture weight, mean vector and covariance matrix for the mixture $i$ from the set of $M$ mixture components. The goal of the speaker identification tasks is to find the index $\hat{S}$ of the speaker from the set of speakers $\mathcal{S}$ whose model will have the maximum a posteriori probability given the observation sequence $O$ :

$$\hat{S} = arg \max_{1 \leq k \leq S} (p(O)|\lambda_k) = arg \max_{1 \leq k \leq S} \prod_{t=1}^{T} p(o(t)|\lambda_k) \quad (7)$$

## 4. EXPERIMENTS AND RESULTS

The lip feature extraction algorithm is tested on audio-visual sequences form the Clemson University database CUAVE, [1]. This database consists of 36 speakers having different skin and lip tones, with various visual occlusions such as: glasses, beards and hats. We consider only part of the database where an individual speaker is present in the scene frontally facing the camera, pronouncing the sequence of digits from "zero" to "nine", five times in repetition. For the majority of the video clips used (33 out of 36), the method performed well. In certain cases the problem appeared either due to the thin lips of the subject and/or significant head movement. For these sequences lip points of interest were marked manually every 10 frames.

After implementing the previously described algorithm, width and height of the lips are extracted from each frame of the video clip and used to form a visual feature vector.

### 4.1. Speaker identification results

Using the obtained visual features we perform visual-only speaker identification experiments. For each trial, the training set consists of the four repetitions of the sequence of digits from "zero" to "nine", while the remaining fifth sequence

is used for testing. The identification results for each test trial are represented in Table 1. The best overall results are obtained with number of Gaussian mixtures between three and five.

| test sequence | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| scores [%] | 77.78 | 88.89 | 94.44 | 94.44 | 83.33 |

**Table 1**. Visual-only speaker identification scores for five test sequences.

Further on, we perform audio-only and audio-visual identification experiments. Acoustic features are represented with 12 Mel Frequency Cepstral Coefficients (MFCCs) commonly used features in speech recognition. Tests were done for both clean and degraded audio conditions. Noisy environment is simulated by adding white Gaussian noise to the speech signal at different SNR levels ranging from 30 dB to $-6$ dB in steps of 6 dB. All models were trained using clean speech and tested on noisy data.

For audio-visual experiments we use the composite audio-visual feature vector i.e. feature fusion strategy. In order to have the same sampling rate for both audio and visual data streams, the audio is downsampled to the rate of video, or equivalently 30 Hz. Identification scores when using joint audio-visual feature vector are given in Table 2 for both clean and noisy conditions. It can be seen that the help of visual modality significantly improves the system performance when compared to audio-only identification results in noisy conditions. However, the overall audio-visual speaker identification system performance is far from satisfying. The lower scores than the video-only are due to the chosen feature fusion strategy and audio and visual streams misalignment.

| | audio | visual | audio-visual |
|---|---|---|---|
| clean | 100 | 83.33 | 100 |
| 30 dB | 91.67 | 83.33 | 97.22 |
| 24 dB | 93.94 | 83.33 | 94.44 |
| 18 dB | 50.00 | 83.33 | 77.78 |
| 12 dB | 19.44 | 83.33 | 61.11 |
| 6 dB | 8.33 | 83.33 | 41.67 |
| 0 dB | 5.56 | 83.33 | 30.56 |
| -6 dB | 2.78 | 83.33 | 27.78 |

**Table 2**. Speaker identification results using feature fusion and test sequence #5.

## 5. CONCLUSION

We present a novel approach for locating the lip corners, as well as the upper and the lower outer lip contour middle points, using color transformation and a modified fuzzy c-means clustering algorithm. The proposed method is tested on 36 sequences from the CUAVE database working well for the majority of tested speakers. The coordinates of four key lip feature points are used to estimate the width and height of the speakers' mouth. Then we perform various audio-visual speaker identification experiments. Our preliminary tests done on the CUAVE database show promising results in terms of utilizing the geometric lip features, for the identification tasks on the data corpus, previously used mainly for audio-visual speech recognition.

Future work would be to test the lip feature extraction method in the presence of video noise, as well as to perform audio-visual speaker identification experiments using other visual cues. Since the audio-visual identification scores when using early integration fusion method are still far from satisfying in a noisy environment, the other possible research direction would be towards utilizing adaptive decision fusion methods.

## 6. REFERENCES


[1] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy, "Moving talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *Eurasip JASP*, vol. 11, pp. 1189–1201, Oct. 2002.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

[3] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. on Multimedia*, vol. 4, no. 1, pp. 23–37, 2002.

[4] D. Dean, P. Lucey, and S. Sridharan, "Audio-visual speaker identification using the CUAVE database," in *Auditory-Visual Speech Processing, AVSP'05*, British Columbia, Canada, July 2005.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR2001*, 2001.

[6] H. J. Trussell, M. J. Vrhel, and E. Saber, "Color image processing [basics and special issue overview]," *IEEE Signal Processing Mag.*, vol. 22, no. 1, pp. 14–22, 2005.

[7] A. W.-C. Liew, S.-H. Leung, and W.-H. Lau, "Segmentation of color lip images by spatial fuzzy clustering," *IEEE Trans. on Fuzzy Systems*, vol. 11, no. 4, pp. 542–549, 2003.

[8] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.